Temporal Consistent Oil Painting Video Stylization

Luyao Zhang, Wenjing Wang, Jiaying Liu*
Wangxuan Institute of Computer Technology, Peking University, Beijing, China

2023 IEEE International Symposium on Circuits and Systems (ISCAS) | 978-1-6654-5109-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCAS46773.2023.10182168

***Abstract*—The automatic rendering of oil painting style video has great artistic and commercial application value. Temporal consistency is the bottleneck of video rendering. However, existing translation methods are either designed for images, or have high training/inference costs on videos due to the estimation of optical flows. This paper explores how to render videos in oil painting styles without video training data. We adopt a motion-based regularization in the training phase and a feature statistics sharing strategy in the inference phase. Experiments show that our model can render vivid and temporally smooth oil painting videos.**

I. Introduction

Oil painting has been one of the most common artistic styles for several centuries, spreading from Europe to the rest of the world. In the past, oil painting requires professional skills. Nowadays, with the development of computer vision, automatic tools for rendering digital photos to oil paintings have emerged [1], [2]. In this paper, we focus on a more difficult task: automatically making oil painting videos. It has important application prospects in many fields, including but not limited to film and television production, video social platforms, advertising media, and art design.

Style transfer is a classical technique for automatic painting. Early methods [3], [4] are based on texture synthesis. In 2015, Gatys *et al*. [1], [5] first used pretrained neural networks [6] to characterize artistic styles. This technique is called neural style transfer (NST). The core of NST is to match feature distributions [7]. Later on, various distribution matching policies have been adopted for style transfer [8]–[10]. Some works also focus on speeding up the style transfer process by building feed-forward frameworks. These methods can be classified into per-style-per-model [11], multiple-style-per-model [12], [13], and arbitrary-style-per-model [14]–[17].

To extend style transfer from image to video, temporal consistency has been introduced into optimization algorithm [18] and network design [19]. Despite the huge success of style transfer, oil painting is still a hard case for state-of-the-art methods. It is because oil painting has vivid stoke textures, rich and dense colours, and a wide range from light to dark. Moreover, many style transfer methods require a reference style image, which is hard to obtain in many cases and improper to ask the user to prepare.

Instead of using style transfer, we render oil painting videos based on generative adversarial networks (GANs), which have better synthesis ability and do not require a reference artwork. Pix2Pix [20] is the first image-to-image translation model, which can learn a mapping between two image domains, *e.g*., a photo and a painting domain. To get rid of the requirement of paired training data, CycleGAN [2] introduces a cycle mapping loop into the training pipeline. Huang *et al*. [21] later proposed multi-modal unsupervised image-to-image translation (MUNIT). Huang *et al*. assumed that the image representation can be decomposed into a content code that is domain-invariant, and a style code that captures domain-specific properties. MUNIT recombines the content code with a random style code sampled from the style space of the target domain. In this way, MUNIT is capable of one-to-many mapping, which can provide more choices for users. Although these GAN models are powerful for rendering oil paintings, they are designed for images and can cause temporal flicking when applied to videos.

To generate sequences that are both temporally smooth and realistic in individual frames, some researchers explore the task of video translation. ReCycleGAN [22] learns a recurrent temporal predictor and introduces a new cycle loss across domains and time. Engelhardt *et al*. [23] used temporal discriminators that take consecutive frames. Recent approaches estimate optical flow to characterize temporal consistency. Vid2Vid [24] combines the optical flow and video-specific constraints. Mocycle-GAN [25] explicitly models the motion across frames in the form of optical flow throughout the translation. Park *et al*. [26] computed flow field to warp the previous output frame onto the current time step. Although optical flow can efficiently characterize the motion across frames, it requires extra computation, which takes a higher cost. Moreover, existing optical flow estimation techniques are not robust enough to handle complex motion and object appearance in real scenes, limiting the practical application of existing video translation models.

To introduce across-frame motion without optical flow, we propose a simple but comprehensive solution for video translation. Our solution consists of two parts: training and inference. For training, we generate artificial transformations that are able to mimic the motion and local distortion of real videos. Based on this artificial transformation, we introduce a regularization that can guide models to maintain temporal consistency even without video training data. For inference, we point out that instance normalization layers introduce inter-frame variation and propose to eliminate the bad effects of instance normalization by sharing the feature statistics among the whole sequence. Through our regularization-based tempo-

* Corresponding Author. This work was supported by the National Natural Science Foundation of China under Contract No.62172020, and a research achievement of Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

(a) Traditional methods

(b) Ours

Fig. 1: Comparison of training guidance in different methods.

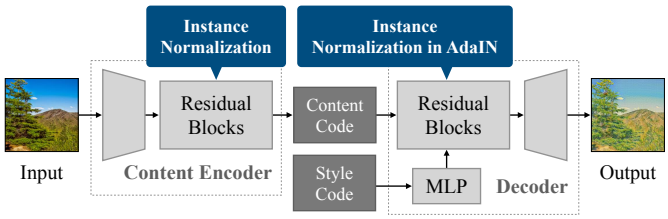


Fig. 2: The inference pipeline of MUNIT [21].

ral consistency learning and sharing normalization strategies, our model can generate vivid and temporally smooth oil painting videos. Experimental results demonstrate the effectiveness of our model. Our strategies can also be extended to other video translation tasks, such as season and flower translation.

The rest of the paper is organized as follows. Sec. II introduces the proposed temporally consistent oil painting video translation framework. Experimental results are shown in Sec. III and concluding remarks are given in Sec. IV.

II. Proposed Method

In this section, we introduce the detailed designs of our oil painting video translation framework.

A. Regularization-based Temporal Consistency Learning

Many video-based methods adopt a temporal loss, which is as follows:

$$||\mathcal{F}(X_t) - \text{Warp}(\mathcal{F}(X_{t-1}), O_{t-1\to t})||, \tag{1}$$

where X_t is the first frame, X_{t-1} is the second frame, and $O_{t-1\to t}$ is the optical flow from X_{t-1} to X_t. $\text{Warp}(X, O)$ represents warping the frame X based on the optical flow O. $\mathcal{F}$ represents the translation network. This temporal loss restricts that when we warp the translated result $\mathcal{F}(X_{t-1})$ from time $t-1$ to time t, it should look similar to the translated result $\mathcal{F}(X_t)$. An example is shown in Fig. 1a.

The problem is that both frames X_{t-1}, X_t, and the corresponding optical flow $O_{t-1\to t}$ need to be provided by the dataset. However, accurate $O_{t-1\to t}$ is hard to obtain as stated in the previous section. In this paper, we propose a training strategy that does not require estimating optical flow.

Given an image X, instead of finding the second frame in the dataset, we synthesize a fake second frame X' by a random optical flow O. An example is shown in Fig. 1b. The benefits are twofold. On the one hand, since the second frame is generated by the optical flow, the optical flow is 100% accurate. On the other hand, since only the "first frame" X needs to be provided by the dataset, we even do not need videos for training, which explicitly improves the convenience of training.

To generate the fake second frame X', we consider two kinds of cross-frame translations. One is object movement, which is represented by a random optical flow O. The other one is a noise δ, such as camera noise and video compression noise. Combine O and δ, we generate the fake second frame as follows:

$$X' = \text{Warp}(X, O) + \delta. \tag{2}$$

Finally, our temporal regularization is as follows:

$$\mathcal{L}_{tmp} = ||\mathcal{F}(X') - \text{Warp}(\mathcal{F}(X), O)||. \tag{3}$$

To implement an oil painting video translation model, we adopt MUNIT [21] as the baseline. MUNIT is an unsupervised image-to-image translation method, which can learn translation without paired data. Moreover, MUNIT can generate multiple results given the same input.

Combining our temporal regularization into the training pipeline of MUNIT, the final training objective is as follows:

$$\begin{aligned}\min_{E_1,E_2,G_1,G_2}\ \max_{D_1,D_2}\ &\mathcal{L}(E_1,E_2,G_1,G_2,D_1,D_2) = \\ &\mathcal{L}^{x_1}_{\text{GAN}} + \mathcal{L}^{x_2}_{\text{GAN}} + \lambda_x(\mathcal{L}^{x_1}_{\text{recon}} + \mathcal{L}^{x_2}_{\text{recon}}) + \\ &\lambda_c(\mathcal{L}^{c_1}_{\text{recon}} + \mathcal{L}^{c_2}_{\text{recon}}) + \lambda_s(\mathcal{L}^{s_1}_{\text{recon}} + \mathcal{L}^{s_2}_{\text{recon}}) + \\ &\lambda_{cyc}(\mathcal{L}^{x_1}_{\text{cyc}} + \mathcal{L}^{x_2}_{\text{cyc}}) + \lambda_t(\mathcal{L}^{x_1}_{tmp} + \mathcal{L}^{x_2}_{tmp}),\end{aligned} \tag{4}$$

where λ_x, λ_c, λ_s, λ_{cyc} and λ_t are weights that balance different loss terms, x_1 represents photographs, and x_2 represents oil painting images. $\mathcal{L}_{\text{GAN}}$ is the adversarial loss that matches the distribution of translated images to the real oil painting distribution. $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{cyc}}$ are bidirectional and cycle reconstruction losses. E_i, G_i, and D_i represent encoders, decoders, and discriminators. Please refer to [21] for their detailed definition.

B. Sharing Normalization

In image-to-image translation models, instance normalization layers are commonly used. For example, the framework of MUNIT in the inference phase is shown in Fig. 2. We can see that there are instance normalization layers in the residual blocks of the content encoder. AdaIN [14] layers in the decoder also have instance normalization. The problem is that instance normalization computes independent statistics (feature mean and variance) for each frame. The inconsistent statistics across frames introduce flicking in video translation, harming temporal consistency.

To solve this problem, we share the statistics for the whole testing video, as illustrated in Fig. 3. We first feed the frames and obtain the mean μ_{F_t} and variance σ_{F_t} in instance normalization layers. Then, we calculate the average mean μ_{avg} and variance σ_{avg}:

$$\mu_{avg} = \frac{1}{T}\sum_{t=1}^{T}\mu_{F_t}, \quad \sigma_{avg} = \frac{1}{T}\sum_{t=1}^{T}\sigma_{F_t}, \tag{5}$$

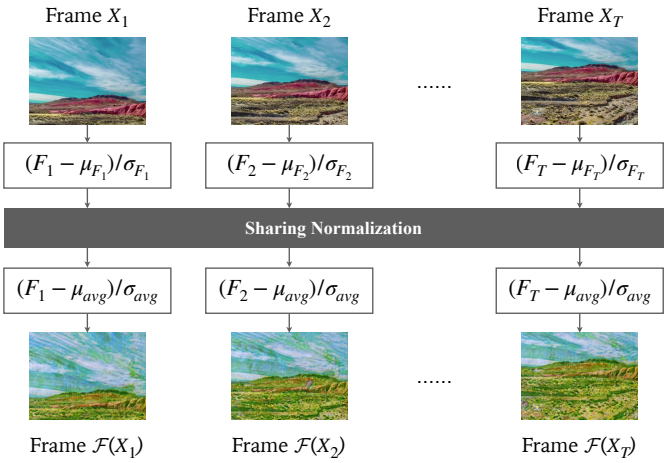


Fig. 3: An illustration for sharing normalization.

(a) Photographs

(b) Oil painting images

Fig. 4: Samples of our video translation training data.

where T is the number of frames in the sequence. Finally, we use the averaged statistics to replace the original frame-wise statistics. Through this strategy, the bad effect of instance normalization layers can be eliminated without extra price.

III. Experiments

A. Experimental Settings

Dataset Preparation. To train the oil painting video translation model, we collect 7038 photographs and 3401 oil painting images from a series of CycleGAN [2] datasets. The training resolution is 256×256. We split 6287 photographs and 2559 oil painting images for training, and 842 photographs and 751 oil painting images for evaluation.

Training Settings. Experiments are conducted on GeForce RTX 2080 Ti and Intel® Xeon® CPU E5-2650 v4 @ 2.20GHz. The training takes 500k iterations. We follow [21] for other parameter settings.

Evaluation Settings. Temporal consistency is evaluated by the widely-used temporal loss [18]. The testing videos are collected from Bilibili[1], a video sharing website. Denote the t-th input frame as X_t, the temporal loss between the t-th and the $(t-1)$-th frame is computed as follows:

$$||M_{t-1\to t} \odot \text{Warp}(\mathcal{F}(X_{t-1}), O_{t-1\to t}) - \mathcal{F}(X_t)||, \quad (6)$$

[1]https://www.bilibili.com/

TABLE I: Temporal loss and FID under different λ_t values.

	λ_t	Temporal loss ↓	FID ↓
Baseline	0	0.0479	**125.10**
Ours	10	0.0416	126.54
	15	0.0312	153.37
	20	**0.0307**	154.35

TABLE II: Performance w/ and w/o sharing normalization.

	Temporal loss ↓	FID ↓
w/o sharing	0.0427	**126.54**
w/ sharing	**0.0416**	**126.54**

where $O_{t-1\to t}$ denotes the estimated ground truth optical flow, which is predicted by PWC-Net [27]. The occlusion mask $M_{t-1\to t}$ is responsible for eliminating the effects of incorrect optical flow predictions and changes in object appearance. Given $O_{t-1\to t}$, $M_{t-1\to t}$ is defined as follows:

$$\nabla_{t-1\to t} = ||\text{Warp}(X_{t-1}, O_{t-1\to t}) - X_t||_1, \quad (7)$$

$$\bar{M}_{t-1\to t} = \tau - \text{CLIP}(\nabla_{t-1\to t}, \tau - 1, \tau), \quad (8)$$

$$M_{t-1\to t} = \bar{M}_{t-1\to t} \times (\mathbf{1} - \text{Warp}(\mathbf{1}, O_{t-1\to t})), \quad (9)$$

where $\mathbf{1}$ is an all-ones matrix, $\text{CLIP}(\cdot, \alpha_{min}, \alpha_{max})$ is a clipping function. We set τ to 10 based on experience. Intuitively, we first characterize incorrect optical flow predictions and object appearance difference by $\bar{M}_{t-1\to t}$, then add movements that are out of space.

To evaluate the synthesis quality, we use Fréchet inception distance (FID) [28] as follows:

$$||\mu_d - \mu_g||_2^2 + tr(\Sigma_d + \Sigma_g - 2\sqrt{\Sigma_d \Sigma_g}), \quad (10)$$

where μ and Σ refer to the mean and covariance matrix of the Inception V3 [29] features. μ_d and Σ_d are of real oil painting images, and μ_g and Σ_g are of generated results. The real oil painting images are from the testing set of our collected data.

B. Ablation Study

We first evaluate the effectiveness of our designs for maintaining temporal consistency.

As shown in Table I, without the proposed temporal consistent regularization (*i.e.*, $\lambda_t = 0$), the model has a temporal loss of about 0.05. With λ_t increasing, the temporal loss decreases, which demonstrates the effectiveness of our motivation. However, λ_t has a negative effect of hurting the synthesis quality. When $\lambda_t > 10$, the FID score becomes more than 150. It is because MUNIT has a limited model ability. Therefore, forcing MUNIT to maintain temporal consistency reduces its capacity to render oil painting styles. To maintain a good balance between temporal consistency and synthesis quality, we choose λ_t = 10 in the following experiments.

Next, we evaluate the effectiveness of sharing the mean and variance of normalization layers. With sharing normalization, the temporal loss decreases from 0.0427 to 0.0416. Since sharing normalization does not change the single-frame synthesis

Fig. 5: Results of oil painting images translation.

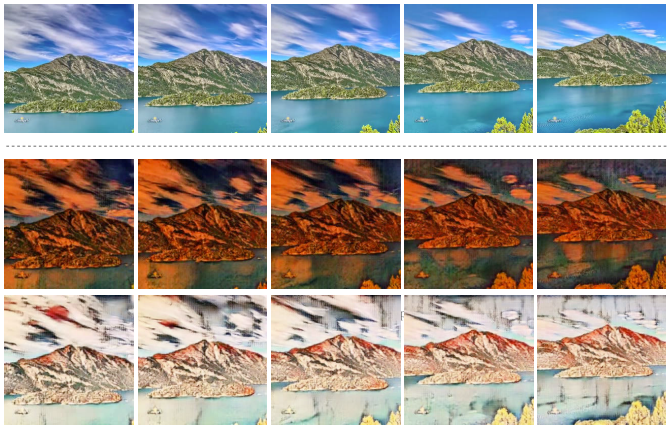


Fig. 6: Results of 256×256 oil painting video translation. The top row is the input and other rows are the outputs.

process, the FID score stays the same. In summary, sharing normalization can improve temporal consistency without hurting the performance of synthesizing oil painting videos.

C. Quantitative Results

Results on images are shown in Fig. 5. Our model can generate vivid oil painting textures, and transfer the color into a classic tone. Moreover, for the same image, our model can generate diverse results, which has high flexibility and usability in practical applications.

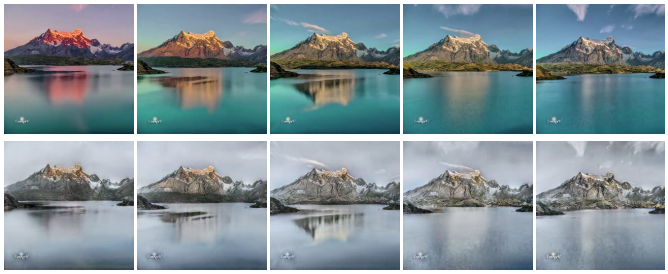
Video oil painting rendering results are shown in Fig. 6. Although the input video has flowing clouds and rivers, our model generates temporally smooth sequences, verifying its robustness in real scenes.

Moreover, despite training on 256×256 resolution, our model can be applied to videos of higher resolution. As shown in Fig. 7, on unseen 512×512 resolution, our model can still render vivid oil painting texture without flickering artifacts, demonstrating the generalization of our method.

D. Application

Besides oil painting, our model can be applied to other video translation tasks as well. We first show results for summer-to-winter season translation in Fig. 8a. Training data is collected from [2]. Although there is a color variation of the sky and mountain in the input sequence, our model is not disturbed by them and generates a temporal smooth video.

Fig. 7: Results of 512×512 oil painting video translation. The top row is the input and other rows are the outputs.



(a) Summer to winter

(b) Dandelion flower to ripe fruits

Fig. 8: Results of other video translation tasks. For each group, the top row is the input and the bottom row is the output.

We further show a hard case of flow translation, which has severe shape deformation. Training data is collected from [22]. As shown in Fig. 8b, our model generates a smooth flower blooming video, demonstrating the robustness of our method.

IV. Conclusion

In this paper, we build an oil painting video stylization framework. We adopt a regularization-based learning strategy and share statistics in the inference phase. Our model can render vivid and temporally smooth oil painting videos.

REFERENCES

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.

[3] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of ACM Special Interest Group on Computer Graphics and Interactive Techniques Conference (SIGGRAPH)*, 2001.

[4] H. Lee, S. Seo, S. Ryoo, and K. Yoon, "Directional texture transfer," in *International Symposium on Non-Photorealistic Animation and Rendering*, 2010.

[5] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *CoRR*, vol. abs/1508.06576, 2015. [Online]. Available: http://arxiv.org/abs/1508.06576

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2015.

[7] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," in *Proceedings of International Joint Conference on Artificial Intelligence, (IJCAI)*, 2017.

[8] C. Li and M. Wand, "Combining markov random fields and convolutional neural networks for image synthesis," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] N. I. Kolkin, J. Salavon, and G. Shakhnarovich, "Style transfer by relaxed optimal transport and self-similarity," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 051–10 060.

[10] N. Kalischek, J. D. Wegner, and K. Schindler, "In the light of feature distributions: moment matching for neural style transfer," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9377–9386.

[11] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[12] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stylebank: An explicit representation for neural image style transfer," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M. Yang, "Diversified texture synthesis with feed-forward networks," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.

[15] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M. Yang, "Universal style transfer via feature transforms," in *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[16] L. Sheng, Z. Lin, J. Shao, and X. Wang, "Avatar-net: Multi-scale zero-shot style transfer by feature decoration," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos," in *Proceedings of German Conference on Pattern Recogition*, ser. Lecture Notes in Computer Science, vol. 9796, 2016, pp. 26–36.

[19] H. Huang, H. Wang, W. Luo, L. Ma, W. Jiang, X. Zhu, Z. Li, and W. Liu, "Real-time neural style transfer for videos," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[20] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] X. Huang, M. Liu, S. J. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.

[22] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-gan: Unsupervised video retargeting," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.

[23] S. Engelhardt, R. D. Simone, P. M. Full, M. Karck, and I. Wolf, "Improving surgical training phantoms by hyperrealism: Deep unpaired image-to-image translation from real surgeries," in *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018.

[24] T. Wang, M. Liu, J. Zhu, N. Yakovenko, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[25] Y. Chen, Y. Pan, T. Yao, X. Tian, and T. Mei, "Mocycle-gan: Unpaired video-to-video translation," in *Proceedings of ACM International Conference on Multimedia (MM)*, 2019.

[26] K. Park, S. Woo, D. Kim, D. Cho, and I. S. Kweon, "Preserving semantic and temporal consistency for unpaired video-to-video translation," in *Proceedings of ACM International Conference on Multimedia (MM)*, 2019.

[27] D. Sun, X. Yang, M. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.